Rapid Data Science Exploration for Cybersecurity
Who Are We?

George
CDS Capsule8
(Now at GSI Tech)

Alex
Data Scientist
Capsule8

Harini
Data Scientist
Capsule8
Outline

1. Cybersecurity Landscape
2. Cybersecurity as a Data Problem
3. Capsule8 Sensor
4. Notebook Demos !
5. Takeaways
Russian Hackers Appear to Shift Focus to U.S. Power Grid

Over 200,000 MikroTik Routers Compromised in Cryptojacking Campaign

Computer virus cripples top Apple supplier TSMC

Reddit Hacked, Some User Data Stolen
CYBER THREAT PREVENTION
Scary Trends

- There is a hacker attack every 39 seconds
- Denial-of-service attacks up 140%
- Cost of breach to $150 million
Opportunities

- Cybersecurity spending tops $100 billion
- Funding for cybersecurity startups up 25%
- Shortfall of 3 million cybersecurity workers
Opportunities

- Cybersecurity spending tops $100 billion
- Funding for cybersecurity startups up 25%
- Shortfall of 3 million cybersecurity workers

Hottest Tech Jobs: Sec DevOps Eng, Data Sci
Rare Skill Set

DATA SCIENCE

CYBER SECURITY
Cybersecurity as a Data Problem

Data Variety
Data Volume
Data Velocity
Data Veracity
Cybersecurity as a Data Problem

Data Variety
Data Volume
Data Velocity
Data Veracity
Production Attacks in the Cloud
# The Technology Stack

<table>
<thead>
<tr>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Libraries</td>
</tr>
<tr>
<td>System Call Interface</td>
</tr>
<tr>
<td>VFS</td>
</tr>
<tr>
<td>File Systems</td>
</tr>
<tr>
<td>Volume Manager</td>
</tr>
<tr>
<td>Block Device Interface</td>
</tr>
<tr>
<td>Device Drivers</td>
</tr>
</tbody>
</table>

Source: [brendangregg.com/perf_events/perf_events_map.png](https://brendangregg.com/perf_events/perf_events_map.png) (creativecommons)
Active Monitoring and Logging

Source: brendangregg.com/perf_events/perf_events_map.png (creativecommons)
Cybersecurity as a Data Problem

Data Variety
Data Volume
Data Velocity
Data Veracity
Big Data Architectures

Log Files → Transform, Collect & Aggregate → Store → Process

- Variety of Formats
- Transform into Common Format
- NOSQL
- Distributed File System
- Batch/Real Time
- hadoop
- Spark
Cybersecurity Vendors

Disclaimer: Vendors shown are representative only. No usage or endorsement should be construed because they are shown here.

Source: RSA Conf
SOC Investigates Alerts 

Security Alerts Dashboard
Alerts

- malware
- invalid logins
- suspicious process
- unusual file activity
- unusual network activity
- privilege escalation
- so many more...

Azure Security Center: 200 types of alerts!
So Many Alerts!

Alert haystack
The majority of bank security chiefs say their staffers sort through more than 100,000 security alerts daily

- 300,000+, 17%
- 200,001 to 300,000, 20%
- 100,001 to 200,000, 24%
- 50,001 to 100,000, 14%
- 10,000 to 50,000, 25%

Source: Ovum
Cybersecurity as a Data Problem

- Data Variety
- Data Volume
- Data Velocity
- Data Veracity
Median Time To Resolve

- Alerts drive response at SOC
- Industry MTTR is 20-30 minutes per alert
- X Humans don’t scale
Target Breach of 2013

Attackers compromised Fazio Mechanical Services.

Attackers broke into Target’s network and tested malware on POS machines.

Attackers began to collect credit card data.

POS malware fully installed.

Attackers installed data exfiltration malware.

Symantec and FireEye alerts triggered.

Attackers began to move credit card data out.

Additional FireEye alerts triggered.

Department of Justice notified Target.

Target removed most malware.

Source: arXiv:1701.04940
Cybersecurity as a Data Problem

- Data Variety
- Data Volume
- Data Velocity
- Data Veracity
Alerts “Veracity”

- More than 50% are false positives
- Cost of FPs at $2 million
- False negatives
Cybersecurity is a Data Challenge!

- Data Variety
- Data Volume
- Data Velocity
- Data Veracity
Opportunities

- Reduce / Eliminate Alerts
- Threat Detection Modeling
- Insider Threats
- Continuous Embeddings
- Adversarial
- APT / Attention Models
Weaponized AI

“The capability to do all of this is real. It exists now.” - Marco Rubio, Senator

“You don’t need software engineers anymore. You just download it to your PC and run it.” - Chris Bregler, Google
Data Science Tools

Data Exploration
Data Prep
Data Validation
Productionalization
Data Science Tools

Data Exploration
Data Prep
Data Validation
Productionalization
# The Technology Stack

<table>
<thead>
<tr>
<th>Device Drivers</th>
<th>System Call Interface</th>
<th>System Libraries</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Device Interface</td>
<td>VFS</td>
<td>Sockets</td>
<td></td>
</tr>
<tr>
<td>Volume Manager</td>
<td>File Systems</td>
<td>Sockets</td>
<td>Scheduler</td>
</tr>
<tr>
<td></td>
<td>TCP/UDP</td>
<td>Sockets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Sockets</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ethernet</td>
<td>Sockets</td>
<td></td>
</tr>
</tbody>
</table>

Source: [brendangregg.com/perf_events/perf_events_map.png](brendangregg.com/perf_events/perf_events_map.png) (creativecommons)
What problem does Capsule8 solve?
What is the Capsule8 open source sensor?

- Enables low level visibility into your OS
- Uses the Linux kernel debugging tools to create telemetry
- Telemetry easily accessible at scale
How much data?

After 8 hours of events

2,000,000+   File
580,000+     Process
658         Container
DEMO!
Takeaways

• Low level visibility into your OS is hard, capsule8 sensors (open source) makes it easy!
• OS Telemetry is the fuel to craft threat detection techniques
• Visualizing security data through Jupyter
Future Work

• Serving visualizations with bokeh servers
• Serving machine learning models
• Productionalizing notebooks for cybersecurity
• Native, security hardened notebooks
https://hub.docker.com/u/cmrfrd/

https://pypi.org/project/capsule8/

https://github.com/capsule8/capsule8

https://gitlab.com/honeycombsecurity/jupytercon2018
Thank you!  Q&A
Contacts

george.williams@gmail.com
@cgeorgewilliams

alexanderjcomerford@gmail.com
@4l3x_c0m3rf0rd

harini.kannan23@gmail.com
@ha_ri_nee
Backup Slides
Demo Deployment